

# UNITARY-GROUP INVARIANT KERNELS AND FEATURES FROM TRANSFORMED UNLABELED DATA

Dipan K. Pal & Marios Savvides

Carnegie Mellon University

Pittsburgh, PA 15213

{dipanp, msavvid}@andrew.cmu.edu

## ABSTRACT

The study of representations invariant to common transformations of the data is important to learning. Most techniques have focused on local approximate invariance implemented within expensive optimization frameworks lacking explicit theoretical guarantees. In this paper, we study kernels that are invariant to the unitary group while having theoretical guarantees in addressing practical issues such as (1) unavailability of transformed versions of *labelled* data and (2) not observing all transformations. We present a theoretically motivated alternate approach to the invariant kernel SVM. Unlike previous approaches to the invariant SVM, the proposed formulation solves both issues mentioned. We also present a kernel extension of a recent technique to extract linear unitary-group invariant features addressing both issues and extend some guarantees regarding invariance and stability. We present experiments on the UCI ML datasets to illustrate and validate our methods.

## 1 INTRODUCTION

It is becoming increasingly important to learn well generalizing representations that are invariant to many common transformations of the data. These transformations can give rise to many ‘degrees of freedom’ even in a constrained task such as face recognition (*e.g.* pose, age-variation, illumination etc.). In fact, explicitly factoring them out leads to improvements in recognition performance as found in Leibo et al. (2014); Hinton (1987). To this end, the study of invariant features is important. Anselmi et al. (2013) showed that features that are explicitly invariant to intra-class transformations allow the sample complexity of the recognition problem to be reduced.

**Prior Art: Invariant Kernels.** Kernel methods in machine learning have long been studied to considerable depth. Nonetheless, the study of invariant kernels and techniques to extract invariant features has received less attention. An invariant kernel allows the kernel product to remain invariant under transformations of the inputs. There has been some work on incorporating invariances into popular machinery such as the SVM in Lauer & Bloch (2008). Most instances of incorporating invariances focused on *local* invariances through regularization and optimization such as Schölkopf et al. (1996; 1998); Decoste & Schölkopf (2002); Zhang et al. (2013). Some other techniques were jittering kernels (Schölkopf & Smola (2002); Decoste & Schölkopf (2002)) and tangent-distance kernels (Haasdonk & Keysers (2002)), both of which sacrificed the positive semi-definite property of its kernels and were computationally expensive. Haasdonk & Burkhardt (2007) had first used group integration to arrive at invariant kernels, however, their approach does not address two important problems that arise in practice (group observation through unlabelled samples and partially observed groups). We will shortly state these problems more concretely and will show that the invariant kernels proposed do in fact solve both problems.

**Prior Art: Invariance through dataset augmentation.** Many approaches in the past have enforced invariance through generating transformed *training* samples in some form such as Poggio & Vetter (1992); Schölkopf & Smola (2002); Schölkopf et al. (1998); Niyogi et al. (1998); Reiser (2008); Haasdonk & Burkhardt (2007). This assumes that one has knowledge about the transformation. The approach presented in this paper however, under the unitarity assumption, can learn the transformations through unlabelled samples and does not need training dataset augmentation. Perhaps, the

most popular method for incorporating invariances in SVMs is the virtual support method (VSV) in Schlkopf et al. (1996), which used sequential runs of SVMs in order to find and augment the support vectors with transformed versions of themselves. Loosli et al. (2007) proposed a similar algorithm to generate and prune out examples. Though these methods have had some success, most of them still lack explicit theoretical guarantees towards invariance. The proposed invariant kernel SVM formulation on the other hand, is guaranteed to be invariant. Further, unlike VSV and other approaches to incorporate invariance into the SVM, the proposed invariant kernel SVM solves the common and important practical problems that we will state shortly. To the best of our knowledge, it is the first formulation to do so.

**Prior Art: Linear Invariant Features.** Recently, Anselmi et al. (2013) proposed linear group-invariant features as an explanation for multiple characteristics of the visual cortex. They achieve invariance in a slightly more general way than group integration, utilizing measures of the distribution characterizing an orbit of a sample under the action group. We extend the method to the RKHS using unitary kernels and extend some properties regarding invariance and stability. We also show that the extension can solve both motivating problems (Problem 1 and Problem 2). This leads to a practical way of extracting non-linear invariant features with theoretical guarantees.

**Motivating Problems.** We now state the two central problems that this paper tries to address through invariant kernels and features. A common practical problem that one faces utilizing previous methods involving generating transformed samples is the computational expense of generating and processing them (including virtual support vectors). Further, in many cases transformed *labelled* samples are unavailable. Two important problems that arise when practically applying invariant kernels and features are:

**Problem 1: (Group observed through unlabelled samples)** *The transformed versions of the training labelled data are not available i.e. one might only have access to transformed versions of unlabelled data outside of the training set (theoretically equivalent to having transformed versions of arbitrary vectors), e.g. only unlabelled transformed images are observed.*

**Problem 2: (Partially observed group)** *Not all members of the group (symmetric set) of transformations  $\mathcal{G}$  are observed i.e. the group is only partially observed through its actions, e.g. not all transformations of an image are observed. In many practical cases, partial invariance is in fact necessary, when a transformation from one class to another exists.*

**Group Theory and Invariance.** Towards this goal, the study of incorporating invariance through group integration seems useful. Group theory is an elegant way to model symmetry. Classical invariant theory provides group integration techniques to enforce invariance. Group integration can also be used to model mean pooling (and max pooling albeit in a different framework as proposed in Anselmi et al. (2013)), which is in implicit use in several areas of machine learning and computer vision. The transformations, in this paper, are modelled as *unitary* and collectively form the unitary-group  $\mathcal{G}$ . Classes of learning problems, such as vision, often have transformations belonging to the unitary-group, that one would like to be invariant towards (such as translation and rotation). The results can also be extended to discrete groups. In practice however, Liao et al. (2013) found that invariance to much more general transformations not captured by this model can be achieved.

We will see that given explicit access to  $\mathcal{G}$ , one can theoretically capitalize on properties such as guaranteed global invariance (as opposed to local invariance in optimization approaches, where the classifier is invariant to only small transformations). However, controlled *local* invariance can also be achieved. Local invariance is important when the extreme transformation of one class overlaps with another. The unitary property of the group and the unitary restriction on kernels (in Section 2.1), allow the development of theoretical motivation for existing techniques, an invariant kernel and invariant kernel features theoretically addressing Problems 1 and 2.

**Contributions.** We list our main contributions below:

1. In contrast to many previous studies on invariant kernels, our focus is to study positive semi-definite unitary-group invariant kernels and features guaranteeing invariance that can address *both* Problem 1 and Problem 2.
2. One of our central results to applying group integration in the RKHS builds on the observation that, under unitary restrictions on the kernel map, group action is preserved in the RKHS.

3. Using the proposed invariant kernel, we present a theoretically motivated alternate approach to designing a non-linear invariant SVM that can handle both Problem 1 and Problem 2 with explicit invariance guarantees.
4. We propose kernel unitary-group invariant feature extraction techniques by extending a theory of linear group invariant features presented in Anselmi et al. (2013). We show that the kernel extension addresses both Problem 1 and Problem 2 and preserves properties such as global (and local) invariance and stability.

**Organization.** The paper is broadly organized into two parts. Section 2 and 3 present the proposed invariant kernels and the invariant kernel SVM, whereas Section 4 and 5 present the proposed invariant features extracted using kernels.

**Section 2 and 3 (Unitary-group Invariant Kernels).** We first present some important known elementary unitary-group integration properties and present a central result to applying group integration in the RKHS in Section 2. We then present a theoretically motivated alternate approach to designing a non-linear invariant SVM and present a simple albeit important result to reduce computation. In Section 3, we then continue on to develop an invariant kernel which does not require to observe transformed versions of the input arguments whatsoever.

**Section 4 and 5 (Unitary-group Invariant Kernel Features).** In Section 4, we propose kernel unitary-group invariant feature extraction techniques by extending a linear invariant feature extraction method (Anselmi et al. (2013)) to the kernel domain. We show that the resultant feature, while addressing Problem 1, preserves important properties such as global invariance and stability. In Section 5, we show that a simple extension of the method can help it to solve both problems (Problem 1 and Problem 2). This leads to a practical way of extracting invariant non-linear features with theoretical guarantees.

**Section 6** presents some experiments illustrating our methods.

## 2 GLOBALLY GROUP INVARIANT KERNELS: WHEN THE GROUP $\mathcal{G}$ IS EXPLICITLY KNOWN

**Premise:** Consider a dataset of normalized samples along with labels  $\mathcal{X} = \{x_i, y_i\} \forall i \in 1 \dots N$  with  $x \in \mathbb{R}^d$  and  $y \in \{+1, -1\}$ . We now introduce into the dataset a number of unitary transformations  $g$  part of the locally compact unitary-group  $\mathcal{G}$  (in general we require local compactness to enable the existence of the Haar measure). Our augmented normalized dataset becomes  $\mathcal{X}_{\mathcal{G}} = \{gx_i, y_i\} \forall g \in \mathcal{G} \forall i^1$ . Thus,  $\mathcal{X} \subseteq \mathcal{X}_{\mathcal{G}}$ . We assume for now that  $\mathcal{G}$  is known and accessible completely. Let  $\phi$  be some mapping to a high dimensional Hilbert space  $\mathcal{H}$ , i.e.  $\phi : \mathcal{X} \rightarrow \mathcal{H}$ . Once the points are mapped, the problem of learning a separator in that space can be assumed to be linear.

An invariant function is defined as follows.

**Definition 2.1 ( $\mathcal{G}$ -Invariant Function).** For any group  $\mathcal{G}$ , we define a function  $f : \mathcal{X} \rightarrow \mathbb{R}^n$  to be  $\mathcal{G}$ -invariant if  $f(x) = f(gx) \forall x \in \mathcal{X} \forall g \in \mathcal{G}$ .

One method of generating an invariant towards a group is through group integration. Group integration has stemmed from classical invariant theory and its foundational theorem was proved by Haar.

**Theorem 2.1. (Haar)** *On every locally compact group there exists at least one left invariant integral. Such an integral is unique except for a strictly positive factor of proportionality.*

One can choose the factor of proportionality such that the group volume equals 1 (i.e.  $\int 1dg = 1$ , in the case of discrete finite groups each group element would be scaled down by  $\frac{1}{|\mathcal{G}|}$ ). For compact groups, such an integral converges for any bounded function of the group. For discrete groups, the integral is replaced by a sum. Group integration can be shown to be a projection onto a  $\mathcal{G}$ -invariant subspace. Such a subspace can be defined for a Hilbert space  $\mathcal{H}$  by  $\mathcal{H}^{\mathcal{G}} = \{x \in \mathcal{H} \mid x = gx \forall g \in \mathcal{G}, \forall x \in \mathcal{H}\}$ . An invariant to any group  $\mathcal{G}$  can be generated through the following basic (previously) known property (Lemma 2.2) based on group integration.

<sup>1</sup>With a slight abuse of notation, we denote by  $gx$  the action of group element  $g \in \mathcal{G}$  on  $x$

**Lemma 2.2.** (*Invariance Property*) Given a vector  $\omega \in \mathbb{R}^d$ , and any group  $\mathcal{G}$ , for any fixed  $g' \in \mathcal{G}$  and a normalized Haar measure  $dg$ , we have  $g' \int_{\mathcal{G}} g \omega dg = \int_{\mathcal{G}} g \omega dg$

The Haar measure ( $dg$ ) exists for every locally compact group and is unique upto a positive multiplicative constant (hence normalized). A similar property holds for discrete groups. The Invariance Property results in global invariance to group  $\mathcal{G}$ . This property allows one to generate a  $\mathcal{G}$ -invariant subspace in the inherent space  $\mathbb{R}^N$ .

The following two lemmas (Lemma 2.3 and 2.4) showcase (novel) elementary properties of the operator  $\Psi = \int_{\mathcal{G}} g dg$  for the unitary-group  $\mathcal{G}$ . These properties would prove useful in the analysis of unitary-group invariant kernels and features.

**Lemma 2.3.** If  $\Psi = \int_{\mathcal{G}} g dg$  for unitary  $\mathcal{G}$ , then  $\Psi^T = \Psi$

**Lemma 2.4.** (*Unitary Projection*) If  $\Psi = \int_{\mathcal{G}} g dg$  for any  $\mathcal{G}$ , then  $\Psi\Psi = \Psi$ , i.e. it is a projection operator. Further, if  $\mathcal{G}$  is unitary, then  $\langle \omega, \Psi\omega' \rangle = \langle \Psi\omega, \omega' \rangle \forall \omega, \omega' \in \mathbb{R}^d$

The proofs of these lemmas utilize elementary properties of groups, invariance of the Haar measure  $dg$  and the unitarity of  $g^2$ .

**Sample Complexity and Generalization.** On applying the operator  $\Psi$  to the dataset  $\mathcal{X}$ , all points in the set  $\{gx \mid g \in \mathcal{G}\}$  for any  $x \in \mathcal{X}$  map to the same point  $\Psi x$  in the  $\mathcal{G}$ -invariant subspace. Theoretically, this would drastically reduce sample complexity while preserving linear feasibility (separability). It is trivial to observe that a perfect linear separator learnt in  $\mathcal{X}_{\Psi} = \{\Psi x \mid x \in \mathcal{X}\}$  would also be a perfect separator for  $\mathcal{X}_{\mathcal{G}}$ , thus in theory achieving perfect generalization. We prove a similar result for the RKHS case in Section 2.2. This property is theoretically powerful since cardinality of  $\mathcal{G}$  can be large. A classifier can avoid having to observe transformed versions  $\{gx\}$  of any  $x$  and yet generalize.

## 2.1 GROUP ACTIONS RECIPROCATATE IN A REPRODUCING KERNEL HILBERT SPACE

Group integration provides exact invariance in the domain of  $\mathcal{X}$ . However, it requires the group structure to be preserved. In the context of kernels, it is imperative that the group relation between the samples  $x \in \mathcal{X}_{\mathcal{G}}$  be preserved in the kernel Hilbert space  $\mathcal{H}$  corresponding to some kernel  $k$ . Under the restriction of unitary  $k$ , this is possible. We present an elementary albeit important result that allows this after defining unitary kernels in the following sense.

**Definition 2.2** (*Unitary Kernel*). We define a kernel  $k(x, y) = \langle \phi(x), \phi(y) \rangle$  to be a unitary kernel if, for a unitary group  $\mathcal{G}$ , the mapping  $\phi(x) : \mathcal{X} \rightarrow \mathcal{H}$  satisfies  $\langle \phi(gx), \phi(gy) \rangle = \langle \phi(x), \phi(y) \rangle \forall g \in \mathcal{G}, \forall x, y \in \mathcal{X}$ .

The unitary condition is fairly general, a common class of unitary kernels is the RBF kernel. We now define an operator  $g_{\mathcal{H}} : \phi(x) \rightarrow \phi(gx) \forall \phi(x) \in \mathcal{H}$  for any  $g \in \mathcal{G}$  where  $\mathcal{G}$  is unitary.  $g_{\mathcal{H}}$  thus is a mapping within the RKHS. Under unitary  $\mathcal{G}$ , we then have the following result.

**Theorem 2.5.** (*Covariance in the RKHS*) If  $k(x, y) = \langle \phi(x), \phi(y) \rangle$  is a unitary kernel in the sense of Definition 2.2, then  $g_{\mathcal{H}}$  is unitary, and the set  $\mathcal{G}_{\mathcal{H}} = \{g_{\mathcal{H}} \mid g_{\mathcal{H}} : \phi(x) \rightarrow \phi(gx) \forall g \in \mathcal{G}\}$  is a unitary-group in  $\mathcal{H}$ .

Theorem 2.5 shows that the unitary-group structure is preserved in the RKHS. This provides new theoretically motivated approaches to achieve invariance in the RKHS. Specifically, a theory of invariance which was proposed to utilize unsupervised linear filters can now also utilize non-linear supervised ‘templates’ as we discuss in Section 4.

## 2.2 INVARIANT NON-LINEAR SVM: AN ALTERNATE APPROACH THROUGH GROUP INTEGRATION

We present the group integration approach to kernel SVMs before comparing it to other methods. The decision function of SVMs can be written in the general form as  $f_{\theta}(x) = \omega^T \phi(x) + b$  for some bias  $b \in \mathbb{R}$  (we agglomerate all parameters of  $f$  in  $\theta$ ) where  $\phi$  is the feature map, i.e.  $\phi : \mathcal{X} \rightarrow \mathcal{H}$ .

<sup>2</sup>All proofs are presented in the supplementary material

Reviewing the SVM, a maximum margin separator is found by minimizing loss functions such as the hinge loss along with a regularizer. In order to invoke invariance, we can now utilize group integration in the the kernel space  $\mathcal{H}$  using Theorem 2.5. All points in the set  $\{gx \in \mathcal{X}_G\}$  get mapped to  $\phi(gx) = g_H\phi(x)$  for a given  $g \in \mathcal{G}$ . Group integration then results in a  $\mathcal{G}$ -invariant subspace within  $\mathcal{H}$  through  $\Psi_{\mathcal{H}} = \int_{\mathcal{G}_H} g_H dg_H$  using Lemma 2.2. Introducing Lagrange multipliers  $\alpha = (\alpha_1, \alpha_2 \dots \alpha_N) \in \mathbb{R}^N$ , the dual formulation (utilizing Lemma 2.3 and Lemma 2.4) then becomes

$$\arg \min_{\alpha} - \sum_i \alpha_i + \frac{1}{2} \sum_i \sum_j y_i y_j \alpha_i \alpha_j \langle \Psi_{\mathcal{H}} \phi(x_i), \Psi_{\mathcal{H}} \phi(x_j) \rangle \quad (1)$$

under the constraints  $\sum_i \alpha_i y_i = 0$ ,  $0 \leq \alpha_i \leq \frac{1}{N} \forall i$ . The separator is then given by  $\omega_{\mathcal{H}}^* = \sum_i y_i \alpha_i \Psi_{\mathcal{H}} \phi(x_i) = \Psi_{\mathcal{H}} \omega^*$  thereby existing in the  $\mathcal{G}_H$ -invariant (or equivalently  $\mathcal{G}$ -invariant) subspace  $\Psi_{\mathcal{H}}$  within  $\mathcal{H}$  (since  $g \rightarrow g_H$  is a bijection). Effectively, the SVM observes samples from  $\mathcal{X}_{\Psi_{\mathcal{H}}} = \{x \mid \phi(x) = \Psi_{\mathcal{H}} \phi(u), \forall u \in \mathcal{X}_G\}$ . If  $\mathcal{G}$  is known, then this provides *exact global* invariance during testing. Further,  $\Psi_{\mathcal{H}} \omega^*$  is a *maximum-margin separator* of  $\{\phi(\mathcal{X}_G)\}$ . This can be shown by the following result.

**Theorem 2.6.** (Generalization) *For a unitary group  $\mathcal{G}$  and unitary kernel  $k(x, y) = \langle \phi(x), \phi(y) \rangle$ , if  $\omega_{\mathcal{H}}^* = \int_{\mathcal{G}_H} g_H dg_H \omega^* = \Psi_{\mathcal{H}} \omega^*$  is a perfect separator for  $\{\Psi_{\mathcal{H}} \phi(\mathcal{X})\} = \{\Psi_{\mathcal{H}} \phi(x) \mid \forall x \in \mathcal{X}\}$ , then  $\Psi_{\mathcal{H}} \omega^*$  is also a perfect separator for  $\{\phi(\mathcal{X}_G)\} = \{\phi(x) \mid x \in \mathcal{X}_G\}$  with the same margin. Further, a max-margin separator of  $\{\Psi_{\mathcal{H}} \phi(\mathcal{X})\}$  is also a max-margin separator of  $\{\phi(\mathcal{X}_G)\}$ .*

The invariant non-linear SVM in objective 1, observes samples in the form of  $\Psi_{\mathcal{H}} \phi(x)$  and obtains a max-margin separator  $\Psi_{\mathcal{H}} \omega^*$ . Theorem 2.6 shows that the margins of  $\phi(\mathcal{X}_G)$  and  $\{\Psi_{\mathcal{H}} \phi(\mathcal{X})\}$  are deeply related and implies that  $\Psi_{\mathcal{H}} \phi(x)$  is a max-margin separator for both datasets. Theoretically, the Invariant non-linear SVM is able to generalize to  $\mathcal{X}_G$  on *just* observing  $\mathcal{X}$  and utilizing prior information in the form of  $\mathcal{G}$  for all unitary kernels  $k$ . This is true *in practice* for linear kernels. For non-linear kernels in practice, however, the invariant SVM still needs to observe and integrate over transformed *training* inputs. We also present the following result for unitary-group invariant kernels which helps in saving computation.

**Lemma 2.7.** (Invariant Projection) *If  $\Psi = \int_{\mathcal{G}} g dg$  for any unitary group  $\mathcal{G}$ , then for any fixed  $g' \in \mathcal{G}$  we have  $\langle \Psi \omega, \Psi \omega' \rangle = \langle g' \omega, \Psi \omega' \rangle \forall \omega, \omega' \in \mathbb{R}^d$*

We provide the proof in the supplementary material. Thus, the kernel in the Invariant SVM formulation can be replaced by the form  $k_{\Psi}(x, y) = \langle \phi(x), \Psi_{\mathcal{H}} \phi(y) \rangle$ , thereby reducing the number of transformed training samples required to be observed by an order of magnitude. *It also allows for the kernel  $k_{\Psi}(x, y)$  to be invariant to the orbit of  $x$ , i.e.  $\{gx\}$  with observing just a single arbitrary point ( $g'x$ ) on the orbit.* Nonetheless, as the formulation stands, it still requires observing the entire orbit of atleast one of the transformed training samples. However, we can get around this fundamental problem as we show in the next section (Section 3).

Note that for the general kernel, the  $\mathcal{G}_H$ -invariant subspace cannot be explicitly computed, it is only implicitly projected upon through  $\Psi_{\mathcal{H}} \phi(x_i) = \int_{\mathcal{G}} \phi(gx_i) dg_H$ . It is important to note that during *testing* however, the SVM formulation will be invariant to transformations of the test sample regardless of a linear or non-linear kernel. Also, interestingly,  $\omega^*$  might be a different decision boundary than  $\omega'^*$  obtained by training the vanilla SVM on  $\mathcal{X}_G$ .

**Positive Semi-Definiteness.** The  $\mathcal{G}$ -invariant kernel map is now of the form  $k_{\Psi}(x, y) = \langle \phi(x), \int_{\mathcal{G}} \phi(gy) dg_H \rangle$ . This preserves the positive semi-definite property of the kernel  $k$  while guaranteeing global invariance to unitary transformations., unlike jittering kernels (Schölkopf & Smola (2002); Decoste & Schölkopf (2002)) and tangent-distance kernels (Haasdonk & Keysers (2002)). If we wish to include invariance to *scaling* however, then we would lose positive-semi-definiteness (it is also not a unitary transform). Nonetheless, Walder & Chapelle (2007) show that conditionally positive definite kernels still exist for transformations including scaling, although we focus of unitary transformations in this paper.

**Partial Invariance.** The invariant kernel SVM formulation (objective 1) also supports partial invariance when  $\mathcal{G}$  is not fully observed (addressing Problem 2), a notion extended to invariant kernel

methods in Section 5. Partial invariance gives one control over the degree of invariance over transformation groups, allowing classes that are transformations of one another (such as MNIST classes 6 and 9) to be discriminated.

**Relating the Virtual Support Vector Method (VSV):** Consider the popular Virtual Support Vector Method (VSV) (Schlkopf et al. (1996)). Here the support vectors are augmented with small (finite) number of transformed versions of themselves. *This assumes that the transformations are explicitly known*, thereby failing to address Problem 1. The augmented training set is used to train another SVM with improved invariance. We show in the following section that the Invariant SVM formulation (objective 1), on the other hand, does address Problem 1. The group integration framework provides a theoretical motivation for the VSV, since at minimum, it suggests having transformed versions of the support vectors. The VSV however, can have different  $\alpha_i$  for different transformed versions of a  $x_i$ , whereas group integration would force them to be the same because the kernel  $k_\Psi(x, y)$  is  $\mathcal{G}$ -invariant. For *linear kernels* we have more benefits. Group integration also suggests building an explicit  $\mathcal{G}$ -invariant subspace before projecting the training set on it. This approach does not increase computation time (for *linear kernels*) while allowing the SVM to generalize to  $\mathcal{G}$ -transformed inputs.

### 3 GLOBALLY GROUP INVARIANT KERNELS: WHEN ACTION OF GROUP $\mathcal{G}$ IS OBSERVED *only* ON *unlabelled* DATA

The previous section introduced a group integration approach to the invariant non-linear SVM. Although the formulation addresses Problem 2, it does not address Problem 1 *i.e.* the kernel  $k_\Psi(x, y) = \langle \int_{\mathcal{G}} \phi(gx) dg_{\mathcal{H}}, \int_{\mathcal{G}} \phi(gy) dg_{\mathcal{H}} \rangle = \langle \Psi_{\mathcal{H}}\phi(x), \Psi_{\mathcal{H}}\phi(y) \rangle$  still requires observing transformed versions of the *labelled* input sample namely  $\{gx \mid gx \in \mathcal{X}_{\mathcal{G}}\}$  (or atleast one of the labelled samples if we utilize Lemma 2.7). We now present an approach to not require the observation of any *labelled* training sample whatsoever.

Assume that for every sample  $x \in \mathcal{X}_{\mathcal{G}}$ , there exists a vector  $u_x$  s.t.  $\phi(x) \approx \phi(T)u_x$ , where  $T$  is an arbitrary unlabelled set (in the form of a column-major matrix) of  $M$  arbitrary templates  $\{t_i\}$  (Note that there exist more informed ways of choosing  $T$ , however to keep the theory general we work with arbitrary template sets). We assume that we have access to transformed versions of each template  $t_i$  *i.e.* we observe  $\mathcal{G}$  *only* through  $\{gt \mid t \in T, g \in \mathcal{G}\}$ . We then have the following result.

**Theorem 3.1.** *For a unitary group  $\mathcal{G}$ , a template set  $T \in \mathbb{R}^{d \times M} = \{t_i\}$  and a unitary kernel  $k(x, y) = \langle \phi(x), \phi(y) \rangle$ , if  $\phi(x) = \phi(T)u_x$  and  $\phi(y) = \phi(T)u_y$ , then the  $\mathcal{G}$ -invariant kernel  $k_\Psi(x, y) = \langle \Psi_{\mathcal{H}}\phi(x), \Psi_{\mathcal{H}}\phi(y) \rangle$  can be written as*

$$k_\Psi(x, y) = \int_{\mathcal{G}} \langle \phi(gT)u_x, \phi(T)u_y \rangle dg_{\mathcal{H}}$$

Theorem 3.1 assumes that the points  $\phi(x)$  lies in the span of  $\phi(T)$ . It allows the kernel  $k_\Psi(x, y)$  to be  $\mathcal{G}$ -invariant for  $x, y$  *i.e.*  $k_\Psi(x, y) = k_\Psi(g'x, g''y) \forall g', g'' \in \mathcal{G}$ . It achieves this while *only* observing transformed versions of the unlabelled template set  $T$ . This is very useful since the use of Theorem 3.1 solves Problem 1 while guaranteeing invariance. Further, in practice, one does not need to have explicit knowledge of the transformations. In many cases, they can simply store the naturally transforming samples (*e.g.* transforming images). A constructed kernel can be applied to any dataset directly provided the same group  $\mathcal{G}$  acts. Coefficients  $u_x$  required for Theorem 3.1 for any  $x \in \mathcal{X}_{\mathcal{G}}$  can be approximated by projecting the sample  $x$  onto the space spanned by  $T$  in the RKHS *i.e.*  $u_x = (\phi(T)^T \phi(T))^{-1} \phi(T)^T \phi(x)$ . This assumes that the kernel matrix  $(\phi(T)^T \phi(T))^{-1}$  is invertible, a condition that can be satisfied by construction.

**Invariant Non-linear SVM through transformed unlabelled data (comparison with the VSV):** The invariant kernel SVM in objective 1 using the invariant kernel  $k_\Psi(x, y)$  achieves invariance through learning the transformation *only* through observed unlabelled data. Further, it does not need multiple runs as opposed to the VSV which requires the generation of transformed labelled examples. Theorem 3.1 allows an invariant kernel to be used *directly* without the computational expense of finding potential support vectors, generating transformations of them and then processing the added samples. Further, invariance helps to reduce sample complexity and improve performance given a number of samples, a phenomenon we observe in our experiments.

#### 4 GLOBALLY GROUP INVARIANT KERNEL FEATURES FROM A SINGLE SAMPLE: WHEN ACTION OF $\mathcal{G}$ IS OBSERVED *only* ON UNLABELLED DATA

Up until now we have studied the properties of the proposed unitary group invariant kernels. We now shift our attention to group invariant *features*. Invariant kernels are a form of an invariant similarity measure and can be used to construct invariant feature maps. Anselmi et al. (2013) proposed linear invariant features that enjoys properties such as global invariance and stability. We extend their method to the RKHS using unitary kernels and extend the invariance and stability properties. We now briefly present their theory of invariance.

##### 4.1 THEORY OF LINEAR INVARIANT FEATURES

Under  $\mathcal{G}$ , the orbit of any sample  $x \in \mathcal{X}$  is defined by  $\mathcal{O}_x = \{gx \mid g \in \mathcal{G}\}$ . As a straightforward albeit elegant observation, *the orbit itself is an invariant under  $\mathcal{G}$ , since  $\mathcal{O}_x = \mathcal{O}_{gx}$* . Measures of such an orbit also provide invariance, such as the high dimensional distribution  $P_x$  induced by the group's action on  $x$ . In fact, Anselmi et al. (2013) show  $P_x$  to be both invariant and unique, *i.e.*  $x \sim x' \Leftrightarrow \mathcal{O}_x \sim \mathcal{O}_{x'} \Leftrightarrow P_x \sim P_{x'}$ , where  $\sim$  denotes membership in the same class. Thus, measures of the distribution, through a finite number of one-dimensional projections  $\{P_{\langle x, t^k \rangle}\}_{k=1}^K$ , can be used as a similarity measure between two orbits<sup>3</sup>. Further, the measures are invariant to the action of unitary group  $\mathcal{G}$ . For unitary group  $\mathcal{G}$ , normalized dot-products and an *arbitrary* template  $t^k$ , an empirical estimate of the 1-dimensional distribution of the projection onto template  $t^k$  can be expressed as  $\mu_n^k(x) = \int_{\mathcal{G}} \eta_n(\langle x, gt^k \rangle) dg$ , where  $\eta_n \forall n \in \{1 \dots N\}$ , a non-linearity, can either estimate the  $n$ -th bin of the CDF or the  $n$ -th moment, the set of which together define  $P_{\langle x, t^k \rangle}$ . In practice, Liao et al. (2013) found that a few or even one of these moments has been shown to be sufficiently invariant. The final signature or feature vector is  $\Delta(x) = (\{\mu_n^1(x)\}, \{\mu_n^2(x)\} \dots \{\mu_n^K(x)\},) \in \mathbb{R}^{NK} \forall n$ .

##### 4.2 GROUP INVARIANT FEATURE EXTRACTION IN KERNEL SPACE FROM A *Single* SAMPLE

We now present a kernel extension of the approach to invariance presented above. We assume access to the set  $\mathcal{U} = \{\mathcal{O}_{t^k} \mid \forall k \in \{1 \dots K\}\}$ , *i.e.* the orbits of  $K$  arbitrary unlabelled vectors or templates. For simplicity, we also assume a compact unitary  $\mathcal{G}$  with finite cardinality  $|\mathcal{G}|$ . Then for every  $g' \in \mathcal{G}$ , we have template  $t_{g'}^k = g't^k$ . Similarly, for unitary kernels (Definition 2.2), the templates in the RKHS behave as transformed versions of each other owing to Theorem 2.5. Therefore,  $t_{\mathcal{H}g'}^k = \phi(t_{g'}^k) = g'_{\mathcal{H}}\phi(t^k) \forall g' \in \mathcal{G}$ . Thus,  $\{g\phi(t^k) \mid g \in \mathcal{G}\}$  form a set of transformed elements for each  $k$  under the action of  $\mathcal{G}$ . Invariance can then be achieved using a form of Equation 7 in Anselmi et al. (2013).

$$\Upsilon_n^k(x) = \frac{1}{|\mathcal{G}|} \sum_g \eta_n(\langle \phi(x), t_{\mathcal{H}g}^k \rangle) \quad (2)$$

$\Upsilon(x)$  can extract non-linear kernel features for any *single* sample  $x \in \mathcal{X}$  that are invariant to the group  $\mathcal{G}$  without ever needing to observe  $\{gx \mid g \in \mathcal{G} \setminus I\}$ <sup>4</sup>. This also solves Problem 1 listed in the introduction. Recall that  $\eta_n$  can either estimate the CDF or the set of moments. In the case of moments, the first moment leads to *mean pooling* and the infinite moment results in *max pooling*. We now show that the kernel feature  $\Upsilon(x)$  continues to satisfy useful properties such as stability (in the Lipschitz sense) *i.e.* a form of a stability result in Anselmi et al. (2013) can be proved using a similar analysis.

**Theorem 4.1.** (Stability) *If  $\Upsilon(x)$  is invariant to a unitary-group  $\mathcal{G}$  and the non-linearities  $\eta_n$  are Lipschitz continuous with constant  $L_{\eta_n}$ , with  $L_{\eta} = \max_n(L_{\eta_n})$  s.t.  $NL_{\eta} \leq \frac{1}{\sqrt{2}}$ , and we have a normalized unitary kernel  $k$  with  $k(x, x) = 1, \forall x \in \mathbb{R}^d$ , then*

$$\|\Upsilon(x) - \Upsilon(x')\|_2^2 < 1 - k(x, x')_H \leq 1 - k(x, x')$$

<sup>3</sup>This follows from Cramer-Wold theorem along with concentration of measures.

<sup>4</sup>Note that even though the features extracted are non-linear, invariance generated is purely towards unitary transformations.

for all  $x, x' \in \mathcal{X}$ . Here  $k(x, x')_H$  is the kernel distance in the Hausdorff sense in  $\mathcal{H}$ , i.e.  $k(x, x')_H = \max_{g, g' \in \mathcal{G}} k(gx, g'x')$ .

A good representation ideally should be stable and the distance between two points in the feature space should be bounded. Unstable representations can skew the feature space and allow for degenerate results. Theorem 4.1 shows that under Lipschitz continuity for the  $\eta_n$  estimation functions and  $k(x, x) = 1, \forall x \in \mathbb{R}^d$ , the kernel feature distance is bounded by the kernel product.

**Discriminative templates:** Equation 2 can be instantiated to extract *discriminative* kernel features, by choosing discriminative instead of arbitrary templates. Let  $\mathcal{U}_{\mathcal{H}} = \phi(\mathcal{U})$ , then for each group element  $g' \in \mathcal{G}$ , one can train  $K$  binary one vs. all classifiers with the  $k^{th}$  template ( $g't^k$ ) labelled as +1 and the rest ( $\{gt^k \mid g \in \mathcal{G} \setminus \{g'\}\}$ ) as -1 for all  $k$ . Recall that the separator  $\omega_{g'}^k$  (for template  $t^k$  as +1 and for group element  $g'$ ) can be expressed  $\omega_{g'}^k = \sum_i \alpha_i y_i \phi(g't^k) = g'_{\mathcal{H}}(\sum_i \alpha_i y_i \phi(t^k)) = g'_{\mathcal{H}} \omega_I^k$  (using Theorem 2.5 and where  $I$  is the identity element of  $\mathcal{G}$ ). Thus,  $\{\omega_I^k, \dots, \omega_{g'}^k\}$  form a set of transformed templates for each  $k$  under the action of  $\mathcal{G}$  using which partial invariance can then be achieved through Equation 2<sup>5</sup>.

## 5 TOWARDS PARTIALLY GROUP INVARIANT KERNELS: WHEN THE GROUP $\mathcal{G}$ IS *partially* OBSERVED THROUGH TRANSFORMED SAMPLES

We extend the notion of partial invariance to the kernel features extracted similar to Equation 2 following the analysis of Anselmi et al. (2013). Partial invariance arises from partially observing the group  $\mathcal{G}$ , i.e. observing only a finite group (may not be a subgroup)  $\mathcal{G}_0 \subseteq \mathcal{G}$ . In practice, this is the most likely case. However, partial invariance can be obtained over the observed subset  $\mathcal{G}_0$  through a local kernel feature, which can also be generalized to locally compact groups. A partially invariant kernel feature is  $\hat{\Upsilon}_n^k(x) = \frac{1}{|\mathcal{G}_0|} \sum_{g \in \mathcal{G}_0} \eta_n(\langle \phi(x), \omega_g^k \rangle)$ .

**Uniqueness:** The analysis for uniqueness in Anselmi et al. (2013) can be applied to  $\hat{\Upsilon}(x)$  with no significant changes, since the group structure is preserved in  $\mathcal{H}$  through Theorem 2.5. In summary, *any two partial orbits with a common point are identical*.

**Invariance:** Theorem 6 from Anselmi et al. (2013) can be applied in  $\mathcal{H}$  with some modification.

**Theorem 5.1. (Partially Invariance)** Let  $\eta_n : \mathbb{R} \rightarrow \mathbb{R}^+$  are a set of bijective and positive functions and  $\mathcal{G}$  be a locally compact group. Further, assuming  $\mathcal{G}_0 \subseteq \mathcal{G}$  and  $\text{supp}(\langle gx, \omega_g^k \rangle) \subseteq \mathcal{G}_0$  (where  $\text{supp}()$  denotes the support), then  $\forall \bar{g} \in \mathcal{G}$  and  $\forall x \in \mathbb{R}^d$ , we have  $\hat{\Upsilon}_n^k(x) = \hat{\Upsilon}_n^k(\bar{g}x)$ .

**Stability:**  $\hat{\Upsilon}(x)$  is stable (in the Lipschitz sense) following the analysis of Theorem 4.1. In particular, we have the following result.

**Theorem 5.2. (Stability of Partially Invariant Feature)** If  $\hat{\Upsilon}(x)$  is partially invariant to the group  $\mathcal{G}_0$  and the non-linearities  $\eta_n$  are Lipschitz continuous with constant  $L_{\eta_n}$ , with  $L_{\eta} = \max_n(L_{\eta_n})$  s.t.  $N L_{\eta} \leq \frac{1}{\sqrt{2}}$ , and we have kernel  $k$  with  $k(x, x) = 1, \forall x \in \mathbb{R}^d$ , then for unitary  $\mathcal{G}$ , if  $\mathcal{G}_0 \subseteq \mathcal{G}$  and assuming  $\text{supp}(\langle gx, \omega_g^k \rangle) \subseteq \mathcal{G}_0$ , then  $\|\hat{\Upsilon}(x) - \hat{\Upsilon}(x')\|_2^2 \leq 1 - k(x, x')$  if  $|\mathcal{G}_0| = 1$ . Further, if  $|\mathcal{G}_0| > 1$  then  $\|\hat{\Upsilon}(x) - \hat{\Upsilon}(x')\|_2^2 < 1 - k(x, x')_H$  for all  $x, x' \in \mathcal{X}$ . Here  $k(x, x')_H$  is the kernel distance in the Hausdorff sense over  $\mathcal{G}_0$  in  $\mathcal{H}$ , i.e.  $k(x, x')_H = \max_{g, g' \in \mathcal{G}_0 \subseteq \mathcal{G}} k(gx, g'x')$ .

Thus  $\hat{\Upsilon}(x)$  can achieve partial invariance provided a limited number of transformations of the unlabelled data. Further, results developed for kernel methods in this section encourage their use in practice since the feature  $\hat{\Upsilon}(x)$  now solves both motivating problems mentioned in Section 1. Note that the notion of and results on partial invariance can be easily applied to the invariant kernel  $k_{\Psi}(x, y)$  proposed in Section 2 and 3 thereby making them practical tools with theoretical guarantees.

<sup>5</sup>Since this is agnostic to the selection of  $\alpha$ , any classifier which can be expressed as a linear combination of the samples in  $\mathcal{H}$  (such as the perceptron, SVM, correlation filters) can be used as discriminative templates to generate invariance.



Table 1: Mean 10-fold cross validation testing classification accuracy (%), of a linear SVM tested on different features of  $\mathcal{X}_{\mathcal{G}_0 T_e}$  while it was trained on the corresponding features of  $\mathcal{X}_{T_r}$ .

Dataset	Raw $\mathcal{X}_{T_e}$	Raw $\mathcal{X}_{\mathcal{G}_0 T_e}$	$\mu_n^k(\mathcal{X}_{\mathcal{G}_0 T_e})$	$\hat{\Upsilon}(\mathcal{X}_{\mathcal{G}_0 T_e})_{RBF}$	$\hat{\Upsilon}(\mathcal{X}_{\mathcal{G}_0 T_e})_{poly}$
banana	55.2	55.2	60.56	<b>61.70</b>	60.37
breast	71.34	64.77	71.56	70.42	<b>71.58</b>
german	76.03	62.48	69.63	<b>69.78</b>	69.63
diabetis	75.67	50.35	65.84	<b>66.20</b>	65.84
image	83.81	57.05	57.62	<b>60.27</b>	57.49
splice	84.54	55.03	55.07	<b>79.83</b>	55.07
thyroid	91.53	52.67	66.76	64.63	<b>68.46</b>
ringnorm	77.26	43.68	<b>57.67</b>	56.48	56.85
twonorm	97.59	31.48	<b>69.21</b>	66.06	64.69
waveform	89.87	63.77	65.50	64.64	<b>66.89</b>

Table 2: Mean 10-fold cross validation testing classification accuracy (%).  $\mathcal{X}_{T_e}$  and S.K  $\mathcal{X}_{\mathcal{G}_0 T_e}$  denote results for a standard kernel and I.K  $\mathcal{X}_{\mathcal{G}_0 T_e}$  denotes it for the invariant kernel (Theorem 3.1).

Dataset	$\mathcal{X}_{T_e}$	S.K $\mathcal{X}_{\mathcal{G}_0 T_e}$	I.K $\mathcal{X}_{\mathcal{G}_0 T_e}$
banana	72.34	47.16	<b>50.67</b>
breast	66.15	62.31	<b>67.27</b>
german	70.60	70.00	70.07
diabetis	73.68	44.67	<b>65.39</b>
image	95.96	43.08	<b>56.34</b>
splice	58.93	55.08	55.08
thyroid	90.48	<b>64.81</b>	64.57
ringnorm	69.16	43.83	<b>50.46</b>
twonorm	97.11	32.66	<b>49.96</b>
waveform	72.86	67.06	67.06

## 6 EXPERIMENTAL VALIDATION

**Goal:** The goal of this section is two fold, to see (1) whether partially invariant kernel features  $\hat{\Upsilon}(x)$  and (2) invariant kernel SVM *i.e.* objective 1 coupled with Theorem 3.1, in practice, are able to address Problem 1 and Problem 2, and (3) whether kernel invariant features offer any advantage over linear invariant features  $\mu(x)$ . We refrain from using discriminative kernel features since our theoretical results does not assume any structure for the templates.

**Set-up and Method:** We use 10 normalized datasets (each with number of samples  $\geq 200$ ) from the UCI ML repository for this task. We form a random 10-fold cross-validation partition (training ( $\mathcal{X}_{T_r}$ )/testing ( $\mathcal{X}_{T_e}$ )) for each dataset  $\mathcal{X}$ . In order to enforce Problem 1, we introduce a number of transformations  $g$  belonging the a randomly chosen set of unitary transformations  $\mathcal{G}_0$  into the test data ( $\mathcal{X}_{T_e}$ ) thereby multiplying the test data size by a factor of  $|\mathcal{G}_0|$ , thus obtaining  $\mathcal{X}_{\mathcal{G}_0 T_e}$  (we set  $|\mathcal{G}_0| = 10$ )<sup>6</sup>. However, we do not augment the training data  $\mathcal{X}_{T_r}$ . We instead generate random vectors or templates  $t \in T$  and augment them using the same unitary transformations  $\mathcal{G}_0$  as the test data (we set  $|T| = 100$  for all experiments). This enforces Problem 1. Problem 2 is inherently enforced to a large degree since it is practically very difficult to generate an entire group. The transformations we introduce are a subset of the unitary group *i.e.*  $\mathcal{G}_0 \subset \mathcal{G}$  with  $|\mathcal{G}_0| = 10$ .

For our first experiment, we compute  $\hat{\Upsilon}(x)$  using the randomly generated transformed templates  $T_{\mathcal{G}_0}$  and use the RBF kernel ( $\sigma = 1$ ) and the polynomial kernel ( $k(x, y) = (\langle x, y \rangle + 1)^d$  with degree  $d = 2$ ). We set  $\eta$  to compute the infinite moment equivalent to max-pooling. As an evaluation, to estimate the separability of the data, we train a linear SVM on the unaugmented (not transformed) data ( $\mathcal{X}_{T_r}$ ) using (1) raw features (Raw baseline), (2) linear invariant features ( $\mu_n^k(x)$  baseline), (3) RBF kernel invariant features ( $\hat{\Upsilon}(x)_{RBF}$ ) and (4) polynomial kernel invariant features ( $\hat{\Upsilon}(x)_{poly}$ ). We then test on the augmented corresponding fold (transformed) of the test data ( $\mathcal{X}_{\mathcal{G}_0 T_e}$ ) after extracting the corresponding feature. We also report the test accuracy of testing on raw  $\mathcal{X}_{T_e}$  as an illustration of the classification difficulty introduced by the transformations added in. The results are summarized

<sup>6</sup>We uniformly set  $|\mathcal{G}_0|$  to a reasonably modest value of 10 in order to keep computational load of multiply-ing the dataset manageable.

in Table 1. For our second experiment, we use the same datasets and generate a random 10-fold partition. Here we always train on the *raw* untransformed ( $\mathcal{X}_{Tr}$ ) fold and test on the raw *transformed* data ( $\mathcal{X}_{GeTe}$ ). We train a standard RBF kernel SVM ( $\sigma = 1$ ) and an invariant SVM using the same kernel as described in Theorem 3.1. We also test the standard kernel SVM on the untransformed data ( $\mathcal{X}_{Te}$ ) as an illustration of the classification difficulty introduced due to transformations. The results are summarized in Table 2.

**Results:** Our first observation is that in almost all of the datasets, even the modestly ( $|\mathcal{G}_0| = 10$ ) added transformations significantly impaired the SVM’s performance (Table 1 and Table 2). Thus we confirm that most of the difficulty in the problem of learning arises from the presence of inherent transformations relating different orbits of the data. Secondly, for both experiments explicitly generating invariance through invariant features (Table 1) and through the invariant kernel (Table 2) helps in performance suggesting that in both cases sample complexity was lowered. We find that invariant kernel features and the invariant kernel (Theorem 3.1), in practice as well, address Problem 1 and Problem 2. Kernel features, in general, modestly outperform linear features in most of these datasets since even though the features are non-linear, the transformation they are invariant to are linear.

## 7 CONCLUSION

One of the main handicaps in applying invariant kernel methods was the computational expense in generating and processing additional transformed form of the data. Further, in many cases it is difficult to generate such samples due to the transformation being unknown. However, in many cases, it is easier to obtain transformed unlabelled samples (such as video sequences in vision). The invariant kernels described in this paper can be used to address such issues while theoretically guaranteeing invariance.

## REFERENCES

- Anselmi, Fabio, Leibo, Joel Z., Rosasco, Lorenzo, Mutch, Jim, Tacchetti, Andrea, and Poggio, Tomaso. Unsupervised learning of invariant representations in hierarchical architectures. *CoRR*, abs/1311.4158, 2013. URL <http://arxiv.org/abs/1311.4158>.
- Decoste, Dennis and Schölkopf, Bernhard. Training invariant support vector machines. *Mach. Learn.*, 46(1-3):161–190, March 2002.
- Haasdonk, B. and Keysers, D. Tangent distance kernels for support vector machines. In *Pattern Recognition, 2002. Proceedings. 16th International Conference on*, volume 2, pp. 864–868 vol.2, 2002.
- Haasdonk, Bernhard and Burkhart, Hans. Invariant kernel functions for pattern analysis and machine learning. In *Machine Learning*, pp. 35–61, 2007.
- Hinton, Geoffrey E. Learning translation invariant recognition in a massively parallel networks. In *PARLE Parallel Architectures and Languages Europe*, pp. 1–13. Springer, 1987.
- Lauer, Fabien and Bloch, Gérard. Incorporating prior knowledge in support vector machines for classification: A review. *Neurocomputing*, 71(7):1578–1594, 2008.
- Leibo, Joel Z, Liao, Qianli, and Poggio, Tomaso. Subtasks of unconstrained face recognition. In *International Joint Conference on Computer Vision, Imaging and Computer Graphics, VISIGRAPP*, 2014.
- Liao, Q., Leibo, J. Z., and Poggio, T. Learning invariant representations and applications to face verification. *Advances in Neural Information Processing Systems (NIPS)*, 2013.
- Loosli, Gaëlle, Canu, Stéphane, and Bottou, Léon. Training Invariant Support Vector Machines using Selective Sampling. In *Large Scale Kernel Machines*, pp. 301–320. MIT Press, Cambridge, MA., 2007.
- Niyogi, P., Girosi, F., and Poggio, T. Incorporating prior information in machine learning by creating virtual examples. In *Proceedings of the IEEE*, pp. 2196–2209, 1998.

- Poggio, T. and Vetter, T. Recognition and structure from one 2d model view: Observations on prototypes, object classes and symmetries. *Laboratory, Massachusetts Institute of Technology*, 1992.
- Reisert, Marco. Group integration techniques in pattern analysis a kernel view. *PhD Thesis*, 2008.
- Schlkopf, B., Simard, P., Smola, A., and Vapnik, V. Prior knowledge in support vector kernels. *Advances in Neural Information Processing Systems (NIPS)*, 1998.
- Schlkopf, Bernhard, Burges, Chris, and Vapnik, Vladimir. Incorporating invariances in support vector learning machines. pp. 47–52. Springer, 1996.
- Schölkopf, Bernhard and Smola, Alexander J. *Learning with kernels: Support vector machines, regularization, optimization, and beyond*. MIT press, 2002.
- Walder, Christian and Chapelle, Olivier. Learning with transformation invariant kernels. In *Advances in Neural Information Processing Systems*, pp. 1561–1568, 2007.
- Zhang, Xinhua, Lee, Wee Sun, and Teh, Yee Whye. Learning with invariance via linear functionals on reproducing kernel hilbert space. In *Advances in Neural Information Processing Systems*, pp. 2031–2039, 2013.

## 8 SUPPLEMENTARY MATERIAL

### 8.1 PROOF OF LEMMA 2.2

*Proof.* We have,

$$g' \int_{\mathcal{G}} g \omega dg = \int_{\mathcal{G}} g' g \omega dg = \int_{\mathcal{G}} g'' \omega dg'' = \int_{\mathcal{G}} g \omega dg$$

Since the normalized Haar measure is invariant, *i.e.*  $dg = dg'$ . Intuitively,  $g'$  simply rearranges the group integral owing to elementary group properties.  $\square$

### 8.2 PROOF OF LEMMA 2.3

*Proof.* We have,

$$\Psi^T = \left( \int_{\mathcal{G}} g dg \right)^T = \int_{\mathcal{G}} g^T dg = \int_{\mathcal{G}} g^{-1} dg^{-1} = \Psi$$

Using the fact  $g \in \mathcal{G} \Rightarrow g^{-1} \in \mathcal{G}$  and  $dg = dg^{-1}$ .  $\square$

### 8.3 PROOF LEMMA 2.4

*Proof.* We have,

$$\Psi \Psi = \int_{\mathcal{G}} \int_{\mathcal{G}} gh dg dh = \int_{\mathcal{G}} \int_{\mathcal{G}} g' dg' dh = \int_{\mathcal{G}} dh \int_{\mathcal{G}} g' dg' = \Psi$$

Since the Haar measure is normalized ( $\int_{\mathcal{G}} dg = 1$ ), and invariant. Also for any  $\omega, \omega' \in \mathbb{R}^d$ , we have  $\langle \omega, \Psi \omega' \rangle = \int_{\mathcal{G}} \langle \omega, g \omega' \rangle dg = \int_{\mathcal{G}} \langle g^{-1} \omega, \omega' \rangle dg^{-1} = \langle \Psi \omega, \omega' \rangle$   $\square$

### 8.4 PROOF OF LEMMA 2.7

*Proof.* We have  $\langle \Psi \omega, \Psi \omega' \rangle = \langle \int_{\mathcal{G}} g \omega, \Psi \omega' \rangle dg = \langle \int_{\mathcal{G}} g' \omega, \Psi \omega' \rangle dg = \langle g' \omega, \Psi \omega' \rangle \int_{\mathcal{G}} dg = \langle g' \omega, \Psi \omega' \rangle$

In the second equality, we fix a group element  $g'$  since the inner-product is invariant using the argument  $\langle \phi \omega, \Psi \omega' \rangle = \langle g' \omega, \Psi \omega' \rangle$ . This is true using Lemma 2.2 and the fact that  $\mathcal{G}$  is unitary. Further, the final equality utilizes the fact that the Haar measure  $dg$  is normalized.  $\square$

### 8.5 PROOF OF THEOREM 2.5

*Proof.* We have  $\langle \phi(gx), \phi(gy) \rangle = \langle \phi(x), \phi(y) \rangle = \langle g_{\mathcal{H}} \phi(x), g_{\mathcal{H}} \phi(y) \rangle$ , since the kernel  $k$  is unitary. Here we define  $g_{\mathcal{H}} \phi(x)$  as the action of  $g_{\mathcal{H}}$  on  $\phi(x)$ . Thus, the mapping  $g_{\mathcal{H}}$  preserves the dot-product in  $\mathcal{H}$  while reciprocating the action of  $g$ . This is one of the requirements of a unitary operator, however  $g_{\mathcal{H}}$  needs to be linear. We note that linearity of  $g_{\mathcal{H}}$  can be derived from the linearity of the inner product and its preservation under  $g_{\mathcal{H}}$  in  $\mathcal{H}$ . Specifically for an arbitrary vector  $p$  and a scalar  $\alpha$ , we have

$$\|\alpha g_{\mathcal{H}} p - g_{\mathcal{H}}(\alpha p)\|^2 = \langle \alpha g_{\mathcal{H}} p - g_{\mathcal{H}}(\alpha p), \alpha g_{\mathcal{H}} p - g_{\mathcal{H}}(\alpha p) \rangle \quad (3)$$

$$= \|\alpha g_{\mathcal{H}} p\|^2 + \|g_{\mathcal{H}}(\alpha p)\|^2 - 2\langle \alpha g_{\mathcal{H}} p, g_{\mathcal{H}}(\alpha p) \rangle \quad (4)$$

$$= |\alpha|^2 \|p\|^2 + \|\alpha p\|^2 - 2\alpha^2 \langle p, p \rangle = 0 \quad (5)$$

$$(6)$$

Similarly for vectors  $p, q$ , we have  $\|g_{\mathcal{H}}(p + q) - (g_{\mathcal{H}} p + g_{\mathcal{H}} q)\|^2 = 0$

We now prove that the set  $\mathcal{G}_{\mathcal{H}}$  is a group. We start with proving the closure property. We have for any fixed  $g_{\mathcal{H}}, g'_{\mathcal{H}} \in \mathcal{G}_{\mathcal{H}}$

$$g_{\mathcal{H}} g'_{\mathcal{H}} \phi(x) = g_{\mathcal{H}} \phi(g' x) = \phi(g g' x) = \phi(g'' x) = g''_{\mathcal{H}} \phi(x)$$

Since  $g'' \in \mathcal{G}$  therefore  $g''_{\mathcal{H}} \in \mathcal{G}_{\mathcal{H}}$  by definition. Also,  $g_{\mathcal{H}} g'_{\mathcal{H}} = g''_{\mathcal{H}}$  and thus closure is established. Associativity, identity and inverse properties can be proved similarly. The set  $\mathcal{G}_{\mathcal{H}} = \{g_{\mathcal{H}} \mid g_{\mathcal{H}} : \phi(x) \rightarrow \phi(gx) \forall g \in \mathcal{G}\}$  is therefore a unitary-group in  $\mathcal{H}$ .  $\square$

## 8.6 PROOF OF THEOREM 2.6

*Proof.* Since  $\Psi_{\mathcal{H}}\omega^*$  is a perfect separator for  $\{\Psi_{\mathcal{H}}\phi(\mathcal{X})\}$ ,  $\exists \rho' > 0$ , s.t.  $\min_i y_i (\Psi_{\mathcal{H}}\phi(x_i))^T (\Psi_{\mathcal{H}}\omega^*) \geq \rho' \forall \{x_i, y_i\} \in \mathcal{X}$ .

Using Lemma 2.4 and Theorem 2.5, we have for any fixed  $g'_{\mathcal{H}} \in \mathcal{G}_{\mathcal{H}}$ ,

$$(\Psi_{\mathcal{H}}\phi(x_i))^T (\Psi_{\mathcal{H}}\omega^*) = (g'_{\mathcal{H}}\phi(x_i))^T (\Psi_{\mathcal{H}}\omega^*)$$

Hence,

$$\min_i y_i (g'_{\mathcal{H}}\phi(x_i))^T (\Psi_{\mathcal{H}}\omega^*) = \min_i y_i (\Psi_{\mathcal{H}}\phi(x_i))^T (\Psi_{\mathcal{H}}\omega^*) \geq \rho' \forall (g'_{\mathcal{H}} \Rightarrow g) \in \mathcal{G}$$

Thus,  $\Psi_{\mathcal{H}}\omega^*$  is perfect separator for  $\{\phi(\mathcal{X}_{\mathcal{G}})\}$  with a margin of at-least  $\rho'$ . It also implies that a max-margin separator of  $\{\Psi_{\mathcal{H}}\phi(\mathcal{X})\}$  is also a max-margin separator of  $\{\phi(\mathcal{X}_{\mathcal{G}})\}$ .  $\square$

## 8.7 PROOF OF THEOREM 3.1

*Proof.* For any fixed  $g'_{\mathcal{H}}$  we find,  $\langle \Psi_{\mathcal{H}}\phi(x), \Psi_{\mathcal{H}}\phi(y) \rangle = \langle g'_{\mathcal{H}}\phi(x), \Psi_{\mathcal{H}}\phi(y) \rangle$  using Lemma 2.4. Choosing  $g'_{\mathcal{H}}$  to be identity and substituting the expansion of  $\Psi_{\mathcal{H}}$ ,  $\phi(x) = \phi(T)u_x$  and  $\phi(y) = \phi(T)u_y$  we have the desired result.  $\square$

## 8.8 PROOF OF THEOREM 4.1

*Proof.* Since  $\eta_n$  are Lipschitz continuous  $\forall n$ , for each  $k$  component of the signature  $\Upsilon_n^k(x)$ , we have

$$\|\Upsilon^k(x) - \Upsilon^k(x')\|_{\mathbb{R}^N}^2 \leq \frac{1}{|\mathcal{G}|^2} \sum_n \left( \sum_{g \in \mathcal{G}_{\mathcal{H}}} L_{\eta_n} |\langle \phi(x), g_{\mathcal{H}}\omega_g^k \rangle - \langle \phi(x'), g_{\mathcal{H}}\omega_g^k \rangle| \right)^2 \quad (7)$$

$$\leq \frac{L_{\eta}^2}{|\mathcal{G}|^2} \sum_n \left( \sum_{g \in \mathcal{G}_{\mathcal{H}}} \|\phi(x) - \phi(x')\|_{\mathcal{H}} \|g_{\mathcal{H}}\omega_g^k\|_{\mathcal{H}} \right)^2 \quad (8)$$

$$\leq N^2 L_{\eta}^2 \|\phi(x) - \phi(x')\|_{\mathcal{H}} \quad (9)$$

where we utilize Cauchy-Schwartz, Theorem 2.5 and the fact that for some  $t^k \in \mathbb{R}^d$ , we have  $\|\omega_g^k\|_2^2 = \|\phi(t^k)\|_2^2 = \langle \phi(t^k), \phi(t^k) \rangle = k(t^k, t^k) = 1$ . Since  $\Upsilon^k(x)$  is invariant to the action of  $\mathcal{G}$  (and consequently  $\mathcal{G}_{\mathcal{H}}$ ),  $\|\Upsilon^k(x) - \Upsilon^k(x')\|_{\mathbb{R}^N} \leq N^2 L_{\eta}^2 \min_{g_{\mathcal{H}}, g'_{\mathcal{H}} \in \mathcal{G}_{\mathcal{H}}} \|g_{\mathcal{H}}\phi(x) - g'_{\mathcal{H}}\phi(x')\|_{\mathcal{H}}^2 = N^2 L_{\eta}^2 \min_{g, g' \in \mathcal{G}} \|\phi(gx) - \phi(g'x')\|_{\mathcal{H}}^2 = N^2 L_{\eta}^2 \|\phi(x) - \phi(x')\|_H^2 = 2N^2 L_{\eta}^2 (1 - k(x, x')_H)$ . If  $N L_{\eta} < \frac{1}{\sqrt{2}}$ , then the map is a contraction and we obtain the desired result by summing over all  $K$  components and dividing by  $K$ .  $\square$

## 8.9 PROOF OF THEOREM 5.1

*Proof.* The proof is very similar to that of Theorem 6 in Anselmi et al. (2013) since Theorem 2.5 allows the group structure of  $\mathcal{G}_{\mathcal{H}}$  to be preserved in  $\mathcal{H}$ .  $\square$

## 8.10 PROOF OF THEOREM 5.2

*Proof.* The first condition follows the exact analysis as in Theorem 4.1. For the second condition to hold, we apply Theorem 5.1 and follow an argument very similar to that of Theorem 4.1.  $\square$